STAT100 Problem Set 2: Exploratory Data Analysis 1

You need to submit a Word document or PDF for this assignment. Make sure you do the following:

1. **Upload only one file/document in ELMS for Problem Set 2.**
2. **Include your name in the document in the upper left-hand corner. Under your name, write STAT 100 and your section number. Write Problem Set 2 centered on the page.**
3. **Number and letter your answers to the questions accordingly.**
4. **Carefully read all problems and follow all instructions.**
5. **Upload the assignment in ELMS before the deadline of Sunday 2/20 at 11:59 PM otherwise it is considered late. Make sure you save your document on your computer or email it to yourself so that you keep an electronic copy.**

STAT100 Problem Sets need to be completed by students in RStudio. Students should refer to the Tutorial for Problem Set 2 as they are working on this problem set. All Tutorials for Problem Sets can be found in the STAT100 ELMS course under Modules.

For this assignment you need to use the *Course_Data_Set.RData* **file,** which includes data collected from students enrolled in introductory statistics courses in various colleges and universities in MD. The population of interest is all students enrolled in introductory statistics courses in the US.:

Problem Set 2 has one question worth 25 points. Read each question carefully and follow all instructions. **Please follow these instructions for providing your responses:**

- **For #1.a., 1.b., and 1.c., you need to provide the R code that you use to generate the output or data display and provide the output as displayed in the RStudio Console or the data display as shown in the Zoom window in RStudio.**
- **For #1.d. and 1.e., you should type your responses directly in the document you submit for Problem Set 2. You should NOT provide R code or RStudio output for #1.d. and 1.e.**

1. (25 points) **Open the** *Course_Data_Set.RData* **file in RStudio.** You need to create data displays for two variables and answer questions related to the distribution of the variables.

    a. Generate two tables (one showing frequencies and one showing percentages) to display the distribution of the *Tattoos* variable AND generate two tables (one showing frequencies and one showing percentages) to display the distribution of the *Age Group* variable. **You need to generate a total of four tables.** In the document you upload for this assignment, include:
        i. the R code you used to generate the four tables, including comments. **IMPORTANT INSTRUCTIONS: you MUST include your name in the comments of the R code for #1.a.**
        ii. an image (or images) of the four tables as shown in the Console of RStudio (NOTE: two tables should show frequencies and two tables should show percentages)

    b. Generate a pie chart to display the distribution of the *Tattoos* variable. In the document you upload for this assignment, include:
        i. the R code you used to generate the pie chart, including any comments
        ii. an image of the pie chart from the Plots/Zoom section of RStudio. **IMPORTANT INSTRUCTIONS: the title of the pie chart MUST include your name.** The title MUST read "Pie Chart of Whether a Student has a Tattoo, created by *FirstName LastName*". For example, if Prof. Griffin created the pie chart, the title would read "Pie Chart of Whether a Student has a Tattoo, created by Matt Griffin".

    c. Generate a bar chart to display the distribution of the *Age Group* variable. In the document you upload for this assignment, include:
        i. the R code you used to generate the bar chart, including any comments
        ii. an image of the bar chart from the Plots/Zoom section of RStudio. **IMPORTANT INSTRUCTIONS: the title of the bar chart MUST include your name.** The title MUST read "Bar Chart of Age Group, created by *FirstName LastName*". For example, if Prof. Griffin created the bar chart, the title would read "Bar Chart of Age Group, created by Matt Griffin".

    d. Using the data summarized in the data displays, what percent of students in the Course Data Set responded that they do not have any tattoos? Answer in a complete sentence.

    e. Using the data summarized in the data displays, in the Course Data Set are there more students in the *29 -35* age group, or more students in the *Over 35* age group? Answer in a complete sentence and indicate the number of students in the *29 -35* age group and in the *Over 35* age group.